

HMM Tagging

PA154 Jazykové modelování (6.2)

Pavel Rychlý

pary@fi.muni.cz

April 6, 2021

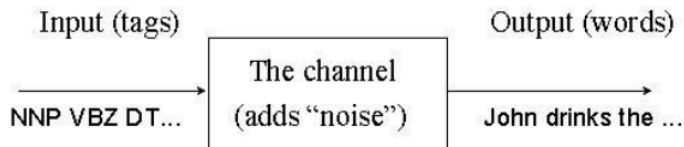
Source: Introduction to Natural Language Processing (600.465)
Jan Hajič, CS Dept., Johns Hopkins Univ.
www.cs.jhu.edu/~hajic

■ Recall:

- ▶ tagging \sim morphological disambiguation
- ▶ tagset $V_T \subset (C_1, C_2, \dots, C_n)$
 - ▶ C_i - morphological categories, such as POS, NUMBER, CASE, PERSON, TENSE, GENDER,...
- ▶ mapping $w \rightarrow \{t \in V_T\}$ exists
 - ▶ restriction of Morphological Analysis: $A^+ \rightarrow 2^{(L, C_1, C_2, \dots, C_n)}$ where A is the language alphabet, L is the set of lemmas
- ▶ extension of punctuation, sentence boundaries (treated as words)

The Setting

- Noisy Channel setting:



- Goal (as usual): discover "input" (T, the tag seq.) given the "output" (W, the word sequence)
 - ▶ $p(T|W) = p(W|T)p(T)/p(W)$
 - ▶ $p(W)$ fixed (W given)... $\text{argmax}_T p(T|W) = \text{argmax}_T p(W|T)p(T)$

The Model

- Two models ($d = |W| = |T|$ word sequence length):
 - ▶ $p(W|T) = \prod_{i=1\dots d} p(w_i | w_1, \dots, w_{i-1}, t_1, \dots, t_d)$
 - ▶ $p(T) = \prod_{i=1\dots d} p(t_i | t_1, \dots, t_{i-1})$
- Too much parameters (as always)
- Approximation using the following assumptions:
 - ▶ words do not depend on the context
 - ▶ tag depends on limited history:
 $p(t_i | t_1, \dots, t_{i-1}) \cong p(t_i | t_{i-n+1}, \dots, t_{i-1})$
 - ▶ n-gram tag "language" model
 - ▶ word depends on tag only: $p(w_i | w_1, \dots, w_{i-1}, t_1, \dots, t_d) \cong p(w_i | t_i)$

The HMM Model Definition

- (Almost) general HMM:
 - ▶ output (words) emitted by states (not arcs)
 - ▶ states: $(n-1)$ -tuples of tags if n -gram tag model used
 - ▶ five-tuple (S, s_0, Y, P_S, P_Y) where:
 - ▶ $S = \{s_0, s_1, \dots, s_T\}$ is the set of states, s_0 is the initial state,
 - ▶ $Y = \{y_1, y_2, \dots, y_Y\}$ is the output alphabet (the words),
 - ▶ $P_S(s_j|s_i)$ is the set of prob. distributions of transitions
 - $P_S(s_j|s_i) = p(t_i|t_{i-n+1}, \dots, t_{i-1}); s_j = (t_{i-n+2}, \dots, t_i), s_i = (t_{i-n+1}, \dots, t_{i-1})$
 - ▶ $P_Y(y_k|s_i)$ is the set of output (emission) probability distributions
 - another simplification: $P_Y(y_k|s_j)$ if s_i and s_j contain the same tag as the rightmost element: $P_Y(y_k|s_i) = p(w_i|t_i)$

■ Use MLE

- ▶ $p(w_i|t_i) = c_{wt}(t_i, w_i)/c_t(t_i)$
- ▶ $p(t_i|t_{i-n+1}, \dots, t_{i-1}, t_i) = c_{tn}(t_{i-n+1}, \dots, t_{i-1}, t_i)/c_{t(n-1)}(t_{i-n+1}, \dots, t_{i-1})$

■ Smooth(both!)

- ▶ $p(w_i|t_i)$: "Add 1" for all possible tag, word pairs using a predefined dictionary (thus some 0 kept!)
- ▶ $p(t_i|t_{i-n+1}, \dots, t_{i-1})$: linear interpolation:
 - ▶ e.g. for trigram model:
$$p'_\lambda(t_i|t_{i-2}, t_{i-1}) = \lambda_3 p(t_i|t_{i-2}, t_{i-1}) + \lambda_2 p(t_i|t_{i-1}) + \lambda_1 p(t_i) + \lambda_0 / |V_T|$$

Unsupervised Learning

- Completely unsupervised learning impossible
 - ▶ at least if we have the tagset given- how would we associate words with tags?
- Assumed (minimal) setting:
 - ▶ tagset known
 - ▶ dictionary/morph. analysis available (providing possible tags for any word)
- Use: Baum-Welch algorithm (see class 15,10/13)
 - ▶ "tying": output (state-emitting only, same dist. from two states with same "final" tag)

- Initialization of Baum-Welch
 - ▶ is some annotated data available, use them
 - ▶ keep 0 for impossible output probabilities
- Beware of:
 - ▶ degradation of accuracy (Baum-Welch criterion: entropy, not accuracy!)
 - ▶ use heldout data for cross-checking
- Supervised almost always better

- "OOV" words (out-of-vocabulary)
 - ▶ we do not have list of possible tags for them
 - ▶ and we certainly have no output probabilities
- Solutions:
 - ▶ try all tags (uniform distribution)
 - ▶ try open-class tags (uniform, unigram distribution)
 - ▶ try to "guess" possible tags (based on suffix/ending) - use different output distribution based on the ending (and/or other factors, such as capitalization)

■ Use Viterbi

- ▶ remember to handle unknown words
- ▶ single-best, n-best possible

■ Another option

- ▶ assign always the best tag at each word, but consider all possibilities for previous tags (no back pointers nor a path-backpass)
- ▶ introduces random errors, implausible sequences, but might get higher accuracy (less secondary errors)

(Tagger) Evaluation

- **A must.** Test data (S), previously unseen (in training)
 - ▶ change test data often if at all possible! ("feedback cheating")
 - ▶ Error-rate based
- Formally:
 - ▶ $\text{Out}(w)$ = set of output "items" for an input "item" w
 - ▶ $\text{True}(w)$ = single correct output (annotation) for w
 - ▶ $\text{Errors}(S) = \sum_{i=1..|S|} \delta (\text{Out}(w_i) \neq \text{True}(w_i))$
 - ▶ $\text{Correct}(S) = \sum_{i=1..|S|} \delta (\text{True}(w_i) \in \text{Out}(w_i))$
 - ▶ $\text{Generated}(S) = \sum_{i=1..|S|} \delta |\text{Out}(w_i)|$

- Accuracy: Single output (tagging: each word gets a single tag)
 - ▶ Error rate: $\text{Err}(S) = \text{Errors}(S)/|S|$
 - ▶ Accuracy: $\text{Acc}(S) = 1 - (\text{Errors}(S)/|S|) = 1 - \text{Err}(S)$
- What if multiple (or no) output?
 - ▶ Recall: $R(S) = \text{Correct}(S)/|S|$
 - ▶ Precision: $P(S) = \text{Correct}(S)/\text{Generated}(S)$
 - ▶ Combination: F measure: $F = 1/(\alpha/P + (1 - \alpha)/R)$
 - ▶ α is a weight given to precision vs. recall; for $\alpha = 0.5$, $F = 2PR/(R + P)$